Multi-Session Group Scenarios for Speech Interface Design[1]

Kari Kanto(1), Maria Cheadle(2), Björn Gambäck(2), Preben Hansen(2), Kristiina Jokinen(1), Heikki Keränen(1), Jyrki Rissanen(1)

(1) Media Lab, UIAH - University of Art and Design Helsinki, Finland
(2) SICS - Swedish Institute of Computer Science AB, Kista, Sweden
{firstname}.{lastname}@uiah.fi, dumas@sics.se

Abstract

When developing adaptive speech-based multilingual interaction systems, we need representative data on the user's behaviour. In this paper we focus on a data collection method pertaining to adaptation in the user's interaction with the system. We describe a multi-session group scenario for Wizard of Oz studies with two novel features: firstly, instead of doing solo sessions with a static mailbox, our test users communicated with each other in a group of six, and secondly, the communication took place over several sessions in a period of five to eight days. The paper discusses our data collection studies using the method, concentrating on the usefulness of the method in terms of naturalness of the interaction and long-term developments.

1 Introduction

In the initial design phase of an adaptive speech-based e-mail application, we needed information on how the users would interact with the envisaged system, and data on vocabulary and language use of the future users. The acquisition of such data poses a problem. Human-human dialogue data is unsuitable, because people speak differently to computers than to each other. Furthermore, real e-mail exchange cannot be monitored because of privacy considerations.

The usual solution is to collect the data in a scenario-driven Wizard of Oz setting, and that was our starting point also. However, this method in its traditional form is insufficient for investigation of long-term temporal phenomena, which are an essential factor when designing adaptive applications. Commonly, a scenario comprises a set of tasks that one test user at a time tries to complete during a single session, and in the e-mail domain, tasks involve a mailbox with several pre-generated messages (e.g., Walker 2000). The sessions usually last a few minutes, rarely exceeding fifteen minutes. In our view, some effects of adaptivity need more time to manifest themselves. For example, the formation of shortcuts, choice of new strategies, and other comparable developments of user expertise may appear only after the user is thoroughly accustomed to the system. A longer period of observation may also illuminate certain linguistic phenomena, such as accommodation and convergence, which are important when dealing with inadequate speech recognition (Zoltan-Ford 1991). Finally, longer stretches of time are needed when studying the complicated interplay between an adaptive system and an adaptive user.

[1] Work sponsored by the European Union's Information Society Technologies Programme under contract IST-2000-29452, DUMAS (www.sics.se/dumas). Thanks to all project participants from KTH and SICS, Sweden; UMIST, UK; ETeX Sprachsynthese AG, Germany; and U. Tampere, U. Art and Design Helsinki, Connexor Oy, and Timehouse Oy, Finland. The wizard interface was developed by Andrew Conroy, UMIST.

Changes in user behaviour over a longer time could perhaps be studied by increasing the number of tasks, but this would soon become tedious for the user and the setting would become increasingly artificial and forced. We have therefore designed a multi-session group scenario method (MSGS) in order to provide a natural setting for observing user behaviour in prolonged system use.

The rest of the paper is laid out as follows: we introduce the MSGS method in Section 2 and the experimental setting in Section 3. In Section 4 we demonstrate the potential of the method with examples from our results. The results are discussed in Section 5, while conclusions are drawn in Section 6.

2 The multi-session group scenario

A Wizard of Oz method with two novel features was conceived: firstly, instead of doing solo sessions with a static mailbox, our test users communicated with each other in groups of six using a simulated speech-based e-mail system via telephone. The e-mail system was controlled by a wizard, and it allowed the subjects to dictate and receive messages, arrange them in folders, etc. Secondly, the communication took place over several sessions in a period of five to eight days during which the subjects played a role defined for them in the scenario. They were advised to call at least twice a day, resulting in a total of around 60 dialogues per experiment.

Having people communicate with each other in the experiment was aimed at creating a more natural and real setting. We wanted the participants to be motivated by the interaction within the group rather than by the interaction with the e-mail system, and to choose their actions based on what happens in the discussion instead of following predestined paths. Methods of dramatic writing (e.g., Egri 1946) were used to prepare a scenario that would incite lively e-mail conversations. Each participant was assigned a fictional character. Five characters in the scenario formed a software development team and the sixth was a customer's representative. All were assigned different tasks and given professional and emotional motives for participating in the e-mail exchange with the others; a tangled web of tensions was woven between them. For example, the customer was the group lead's ex-wife, and the best friend of one of the characters had lost a position in the group to a less competent person, namely the group lead's son. A description of the character assigned to a participant and that character's attitudes towards the others in the group along with some background information and task to solve was presented to each participant at the beginning of the experiment.

The experiments were designed to last for several days in order to gain information on user accommodation and the effects of system adaptivity. The focus was on the development of user expertise and linguistic accommodation.

3 Experimental setup

The six person experiment was duplicated at two sites, UIAH and SICS. The first experiment was conducted in Finnish and the second in Swedish. All participants use computers daily and were 20-35 years old. At both sites three were female and three male. Four participants had previous experience with speech applications, incl. two who were advanced users. One of the participants at UIAH was blind.

The participants were instructed that the speech-based system would basically have the same functionality as an ordinary e-mail program. The participants at SICS were also given an ordinary e-mail account with a web-interface to enable them to review and compose messages in a text modality as well. The users could call from 9 am to 5 pm every day during the experiment.

The telephone calls from the participants were directed through a Dialogic telephony board into a computer where it was recorded and also sent to the loud speakers for the wizards to monitor the call. A wizard interface was used to send the appropriate prompts to the user.

4 Results

To illustrate the potential of the method, we give here examples of different aspects of our results. Generally speaking, the scenario generated active e-mail traffic, even heated discussions, and the subjects seemed committed to playing their characters. A summary of the basic statistics of the two experiments is provided in Table 1.

Table 1: Statistical summary of the experiments

	Finnish participants						
	f_1	f_2	f_3	f_4	f_5	f_6	*average*
number of calls (dialogues)	10	9	16	15	7	4	10.2
messages sent	14	17	19	26	9	3	14.7
messages received	22	19	18	26	11	7	17.2
average call duration (min:sec)	4:28	5:47	5:45	5:24	4:43	4:00	5:01
	Swedish participants						
	s_1	s_2	s_3	s_4	s_5	s_6	*average*
number of calls (dialogues)	9	10	10	16	11	8	10.7
messages sent	18	10	14	9	14	6	11.8
messages received	24	23	26	15	16	16	20.0
average call duration (min:sec)	8:48	5:59	8:04	4:54	3:10	4:32	5:49

The central goal of the method was to study phenomena that occur when a system is used for a prolonged period of time, primarily linguistic developments and changes in the users' level of expertise and in their strategies.

As regards user expertise, one of the participants had a clear curve of development: at the beginning she was hesitant, cancelled actions and didn't always know what to say. After five sessions, she used the system with a highly standardized range of expressions with no signs of hesitation. Another user, who had previous experience with speech systems, picked up the system's vocabulary in her first session and maintained it through the test without exception.

Two users showed other signs of development, by initially restricting themselves to rather short verb phrases when issuing a read-message command, but over time introducing more detailed and specific expansions of the original verb phrase by adding a variety of noun phrases referring to particular messages (Table 2).

Table 2: Developments in phrasing in the Swedish experiment

Initially	After a few sessions
lyssna *listen*	lyssna på det senaste meddelandet från K. *listen to the latest message from K.*
lyssna på meddelandet *listen to the message*	läs båda meddelandena *read both messages*
läs upp *read*	läs upp det första nya meddelandet *read the first new message*
läs upp meddelandet *read the message*	

The shorter phrases were not abandoned but used alongside the more detailed ones throughout the rest of the experiment. The other commands (and some other users, too) show a similar trend, but it is most easily observed in the read-message commands since they are very frequent and were used by all.

Individual interaction strategies were formed, exemplified by one of the participants who developed a habit of calling in once and listening to the new messages, and then calling again a quarter of an hour later to dictate her replies. This pattern becomes evident between her 6th and 7th calls and remains unchanged for the next 8 calls, until the end of the experiment. In her reply to the post-test questionnaire, she mentioned finding it difficult to keep in mind several received messages for reply. Hence, the strategy of separate calls for listening and dictating. This is the sort of development that would not show up in a single-session study.

Another user always accommodated her answers to the system prompts' syntactic form (which becomes easily evident in Finnish as an agglutinative language). For example, when the system asked, "Kenelle haluat lähettää viestisi?" ("To whom do you want to send the message", with the word 'who' in allative case), an accommodating user replied "Katalle" (the name of one of the characters + allative case suffix), whereas an unaccommodating user replied, "Kata" (no case suffix). Some test subjects accommodated only sometimes, usually in sudden situations, and some had no clear tendency towards either.

One of the users, who had previous experience with speech recognisers, never formulated her replies after the model given by the preceding system prompt. She was consistent without exception: for example, when starting to listen to new messages, she said "lue viestit" ("read the messages", plural form) regardless of whether there were many messages or just one.

5 Discussion and future work

The MSGS method seems flexible enough to support observation of several kinds of long-term phenomena. When wizard behaviour is not strictly regulated, the method serves exploration by giving room for unexpected user behaviour. With more strict rules of conduct, the focus moves towards details, for example, testing of particular adaptive functionality. If the experimental system is non-adaptive, users' adaptive behaviour can be studied; otherwise, the interplay between the adaptive system and the adaptive users moves into the forefront.

It is difficult to measure the naturalness and realism of the setting, and presently we must rely on our impressions and the questionnaire that was sent to the test subjects after the experiment. Most participants maintained that they had absorbed or somewhat absorbed their roles, one mentioned that he had not been taken by the scenario at all. All claimed to have been keen to hear what kind of messages they had received. All agreed that they sent fewer and shorter messages in the experiment than they would have if the system had been in real use and the mailbox their own.

The brevity of the messages compared to written e-mails was anticipated due to the modality of the interaction. The fact that we were not able to provide the characters with complete life-long background information is a factor. The relatively small amount of messages, compared to what many users receive in a comparable period of time, is at least partly explained by the small circle of conversants. The researchers "sent" some messages from outside the group and inserted mass mailings, but real-life networks comprise legions of people sending e-mail to each other, which is difficult to simulate.

The rather freeform nature of the setting limits which quantitative measurements can be made. For example, the Dialogue Efficiency Metrics and the Task Success Metrics of the PARADISE evaluation framework (Walker, Litman, Kamm & Abella 1997) are not applicable, because there are no clear-cut tasks for which task success, completion time, etc., could be measured. Dialogue Quality Metrics and User Satisfaction, on the other hand, apply if they are converted to handle

frequencies over several sessions instead of single session packages. This increases the importance of the user questionnaire conducted after the test. Comparisons that require Dialogue Efficiency and Task Success Metrics are better served by the traditional single session method.

As mentioned above some users did not accommodate their replies to the system prompts' syntactic features. This occurred most visibly when the user reacted to clearly defined and simple system prompts when preparing to dictate a new message. The system asked first to whom the user wanted to send the message; answering this question without consideration of the question formulation could mean that the user regarded the situation as slot-filling, in a sort of a transfer from graphical e-mail interfaces. This may open some interesting viewpoints to the mental models the users construct when interacting with speech interfaces.

The vocabulary development of some users, establishing more detailed and specific phrasings of certain commands over time, could be attributed to a number of explanations; the participants developing a better understanding of what the simulated system was able to do, or growing more confident and thus, rather than accepting the system's manner of presenting messages, taking charge of the interaction and making sure it delivered what they wanted.

In future studies, we intend to insert a speech recogniser between the user and the wizard. We hope to create better data by replacing as many as possible simulational parts with real ones without forsaking the quick prototyping possibilities of WOZ setups. The data gathered in the Finnish experiment has been used as a basis for a machine learning experiment (Jokinen, Rissanen, Keränen & Kanto 2002).

6 Conclusions

The paper has described a multi-session group scenario as a means to investigate the effects of system adaptivity on user behaviour and also to provide a realistic simulated environment for testing and developing speech applications.

The rather freeform nature of the setting limits which quantitative measurements can be made: there are no clear-cut tasks for which task success, completion time, etc., could be measured. The simulation provided by the scenario is nonetheless reasonably realistic, giving the designer qualitative insights into the research matter. The effects of adaptivity manifest themselves with the method. Functionality requirements can be explored and dialogue/language models extracted with our scenario. Our conclusion is that the MSGS method is useful and will be developed further.

The scenario can also be used with a working prototype, the system need not be WOZ-operated. This is important in the e-mail domain, since the problem of e-mail privacy remains through the whole system development process.

References

Egri, L. (1946). The Art of Dramatic Writing. London: Simon & Schuster, A Touchstone Book.

Jokinen, K., Rissanen, J., Keränen, H. & Kanto K. (2002). Learning interaction patterns for adaptive user interfaces. *Proc. 7th ERCIM Workshop User Interfaces for All*, Paris, France.

Walker, M.A. (2000). An application of Reinforcement Learning to dialogue strategy selection in a spoken dialogue system for email. *J. Artificial Intelligence Research* **12**, pp. 387-416.

Walker, M.A., Litman, D., Kamm, C., Abella, A. (1997). PARADISE: A Framework for Evaluating Spoken Dialogue Agents. *Proc. 35th Annual Meeting of the Association for Computational Linguistics (ACL-97)*, pp. 271-280, Madrid, Spain, July.

Zoltan-Ford, E. (1991). How to get people to say and type what computers can understand. *Int. J. Man-Machine Studies* **34**, pp. 527-547.